DATA MINING AND KNOWLEDGE DISCOVERY in DATABASE (KDD)

Data Mining vs. KDD

- **Knowledge Discovery in Databases (KDD):** the process of automatic discovery of previously unknown patterns, rules, and other regular contents implicitly present in large volumes of data.
- **Data mining is one of the tasks in the process of knowledge discovery from the database. It is the Use of algorithms to extract the information and patterns derived by the KDD process.**
- *Data Mining* (*DM*) denotes discovery of patterns in a data set previously prepared in a specific way.
- DM is often used as a synonym for KDD.
- However, strictly speaking, DM is just a central phase of the entire process of KDD.

Knowledge discovery

Process of discovering valuable information from a collection of data, or it is the process of converting raw data into useful information

Knowledge discovery is an activity that produces knowledge by discovering it or deriving it from existing information

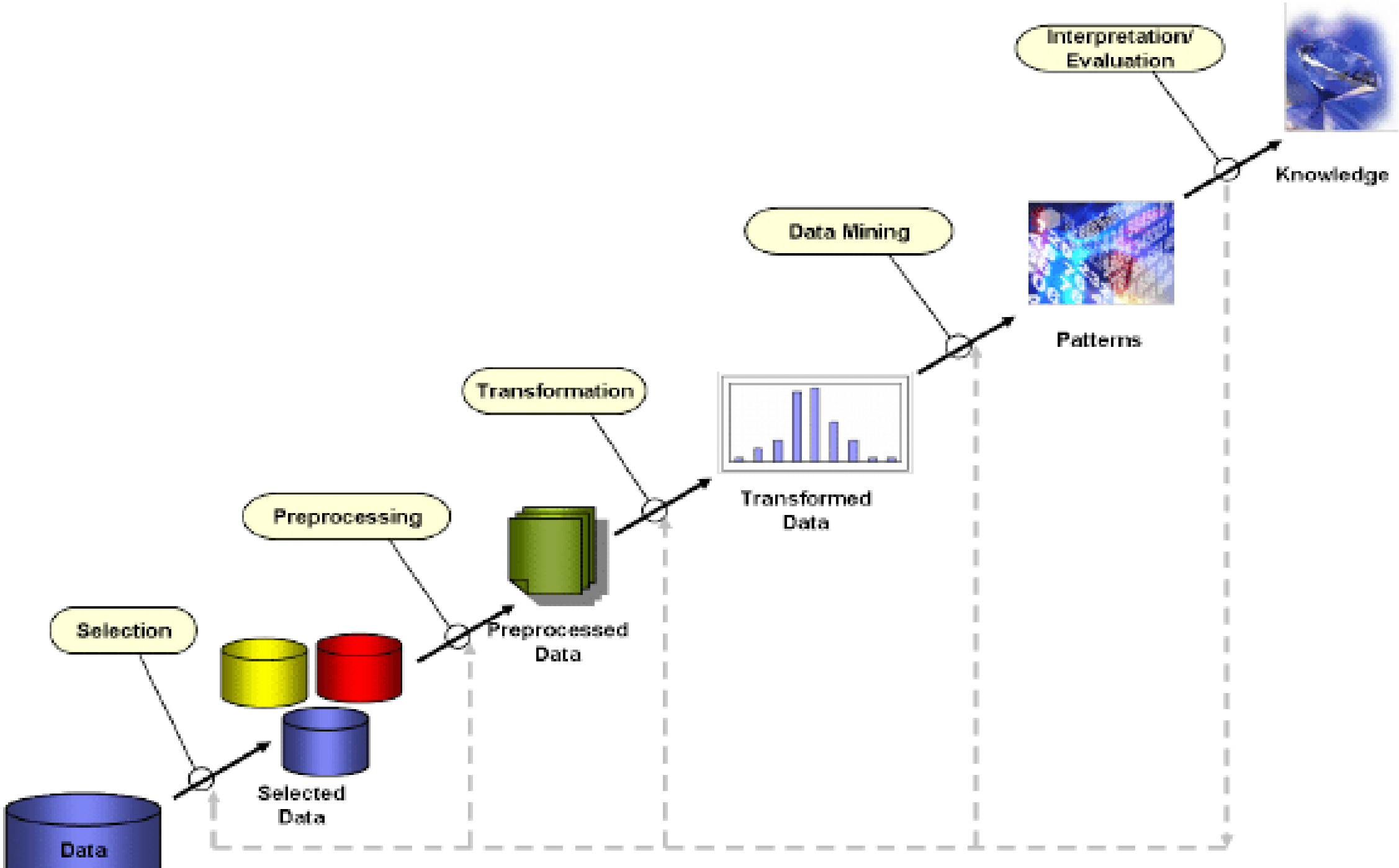
Knowledge discovering refers to the overall process of discovering useful knowledge from data.

Need for knowledge discovery

In the era of information age, there is a lot of data created on day to day basis.

There is need for a new generation of computational theories and tools to assist humans in extracting useful form the rapid growing volumes of digital data

Used in various files such as science ,business like marketing, investment, fraud detection ,telecommunication


Interpretation/
Evaluation
Knowledge
Data Mining
Patterns
Transformation
Transformed
Data
Preprocessing
Preprocessed
Data
Selection
Selected
Data
Data

KDD process……Data Integration

Selection: Obtain data from various sources.

Data Integration is a **data preprocessing technique** that involves combining data from multiple heterogeneous data sources into a coherent data store and provide a unified view of the data.

These sources may include multiple data cubes, databases, or flat files.

KDD process……Data Integration

Data Integration is a **data preprocessing technique** that involves combining data from multiple heterogeneous data sources into a coherent data store and provide a unified view of the data.

These sources may include multiple data cubes, databases, or flat files.

KDD process..... data cleaning and preprocessing

Data cleaning and preprocessing.

This implies eliminating variables or attributes with missing data or eliminating information not useful for this type of task such as text, images, and others.

- Removal of noise or outliers.
- Collecting necessary information to model or account for noise.
- handling missing data fields.
- Accounting for time sequence information and known changes.

KDD process…… data cleaning and preprocessing

Data cleaning is an important step in the data mining process, as it helps to ensure that the data is of high quality and can be easily mined. Data cleaning involves the process of identifying and removing errors, inconsistencies, and missing values in the data.

Some common data cleaning techniques include:

Outlier detection: Identifying and removing data points that are extreme or unusual.

Duplicate detection: Identifying and removing duplicate records from the data.

Missing value imputation: Substituting missing values in the data with statistical estimates or by using machine learning algorithms.

Data transformation: Converting data into a format that can be more easily mined, such as converting categorical variables into numerical variables.

Data validation: Checking the data against a set of validation rules to ensure that it is accurate and complete.

KDD process…… data cleaning and preprocessing

Data standardization: Changing the format of the data to a standard format, such as converting date and time fields to a standard format.

Data reconciliation: Resolving discrepancies in the data by comparing it with other sources of information.

Data scrubbing: Removing sensitive or personal information from the data to protect privacy.

Data sanitization: Removing or masking any sensitive data, such as credit card numbers or social security numbers, to protect privacy.

Data archiving: Storing the data in an archive for long-term retention or for future use.

These techniques help to make sure that the data is accurate, complete, and consistent before it is used for data mining, which can improve the quality of the results.

KDD process…… Transformation

Transformation: Convert to common format. Transform to new format.

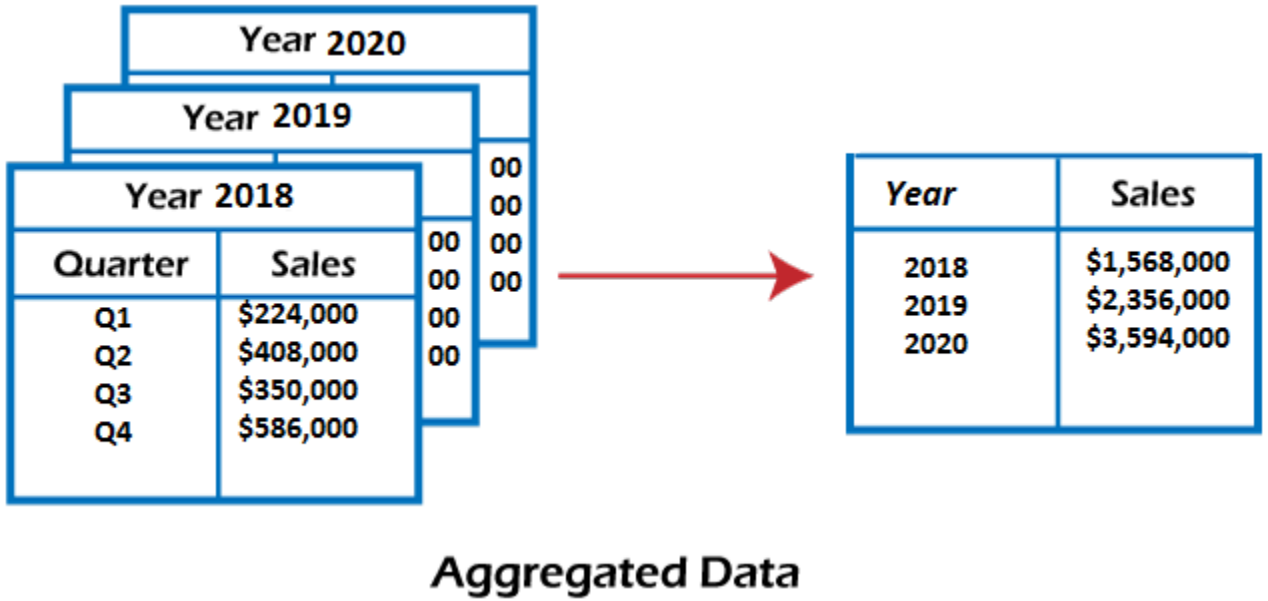
Data transformation is **the process of converting data from one format to another**, typically from the format of a source system into the required format of a destination system.

Data transformation is a component of most data integration and data management tasks, in data warehousing.

The data are transformed in ways that are ideal for mining the data. The data transformation involves steps that are:

1. Smoothing:

It is a process that is used to remove noise from the dataset using some algorithms It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.



2. Aggregation:

Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description.

For **example**, Sales, data may be aggregated to compute monthly& annual total amounts.

KDD process..... Transformation

3. Discretization:
It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.

Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.

For **example**, (1-10, 11-20) (age:- young, middle age, senior).

KDD process..... Transformation

4. Attribute Construction:
Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient.

For example, suppose we have a data set referring to measurements of different plots, i.e., we may have the height and width of each plot.

So here, we can construct a new attribute 'area' from attributes 'height' and 'weight'.

5. Generalization:
It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old).

For **example**, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

6. Normalization: Data normalization involves converting all data variable into a given range.

Min-max,zero-score,decimal scaling,logarithimic,root normalization techniques

Normalization

There are several different normalization techniques that can be used in data mining, including:

Min-Max normalization: This technique scales the values of a feature to a range between 0 and 1. This is done by subtracting the minimum value of the feature from each value, and then dividing by the range of the feature.

where is the current value of feature *F*.

$$v' = \frac{v - \text{minF}}{\text{maxF} - \text{minF}} (new\_max_F - new\_min_F) + new\_min_F \ ,$$

Assume that the minimum and maximum values for the feature F are 50,000 and 100,000 correspondingly. It needs to range *F* from 0 to 1. In accordance with min-max normalization, *v* = 80,000 is transformed to:

$$v' = \frac{80{,}000 - 50{,}000}{100{,}000 - 50{,}000} + (1 - 0) + 0 = \frac{3}{5} = 0{,}6$$

Z-score normalization/zero-mean normalization: This technique scales the values of a feature to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the feature from each value, and then dividing by the standard deviation.

$$v' = \frac{v - \bar{F}}{\sigma_F},$$

Here is the mean and is the standard deviation of feature *F*.

Assume the mean of feature is 65,000 and its standard deviation is 18,000. Applying the z-score normalization we get the following mean of the value equals to 85,800:

$$\frac{85{,}800 - 65{,}000}{18{,}000} = 1{,}156.$$

Decimal Scaling: This technique scales the values of a feature by dividing the values of a feature by a power of 10.

$$v' = \frac{v}{10^j}.$$

Let the input data is: -10, 201, 301, -401, 501, 601, 701 To normalize the above data

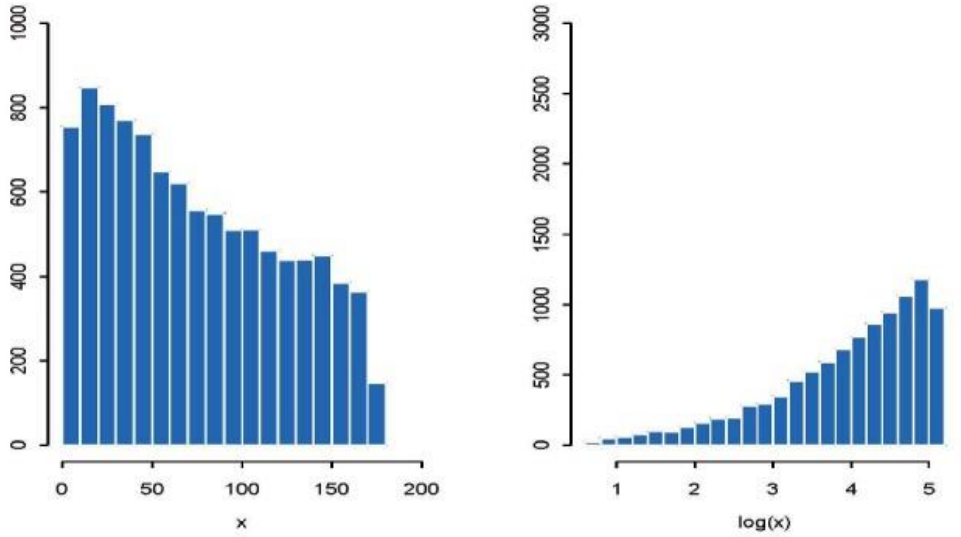
Maximum absolute value in given data(m): 701

Divide the given data by 1000 (i.e j=3)

The normalized data is: -0.01, 0.201, 0.301, -0.401, 0.501, 0.601, 0.701

Logarithmic transformation: This technique applies a logarithmic transformation to the values of a feature. This can be useful for data with a wide range of values, as it can help to reduce the impact of outliers.

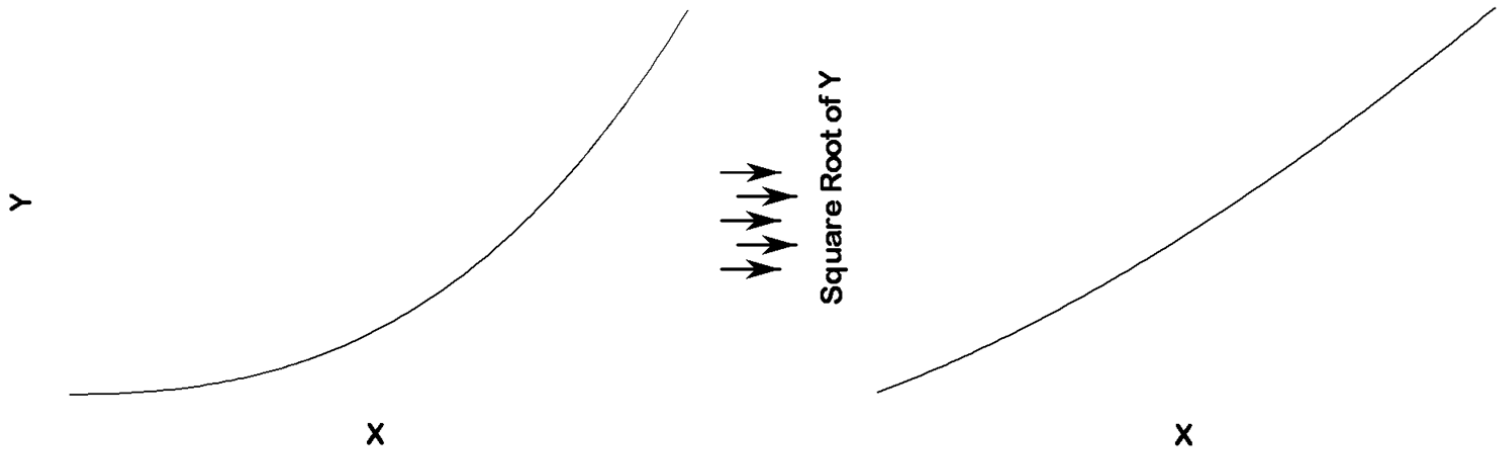
data transformation method in which it replaces each variable x with a log(x).



Histograms of original data (left plot) and log-transformed data (right plot) from a simulation study that examines the effect of log-transformation on reducing skewness.

Root transformation: This technique applies a square root transformation to the values of a feature. This can be useful for data with a wide range of values, as it can help to reduce the impact of outliers.

One solution to fix a non-linear relationship between X and Y, is to try a log or square root transformation.



After the transformation, the relationship looks linear enough to run a linear regression

It’s important to note that normalization should be applied only to the input features, not the target variable, and that different normalization technique may work better for different types of data and models.

KDD process……Data mining

In this step, intelligent methods are applied in order to extract data patterns.

Data mining techniques can be categorized as:

Classification: is the task of generalizing known structure to apply to new data.

Regression: is used to map a data item to a real valued prediction variable.

Clustering: is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data

Association Rules learning : Searches for relationships between variables

Anomaly detection: The identification of unusual data records, that might be interesting or data errors that require further investigation

Summarization: providing a more compact representation of the data set, including **visualization** and **report** generation.

KDD process……interpretation

- **Interpretation**: Present results to user in meaningful manner.

Knowledge representation: knowledge is represented by various visualize tools

Table

Chart

graph

CRISP-DM PROCESS MODEL

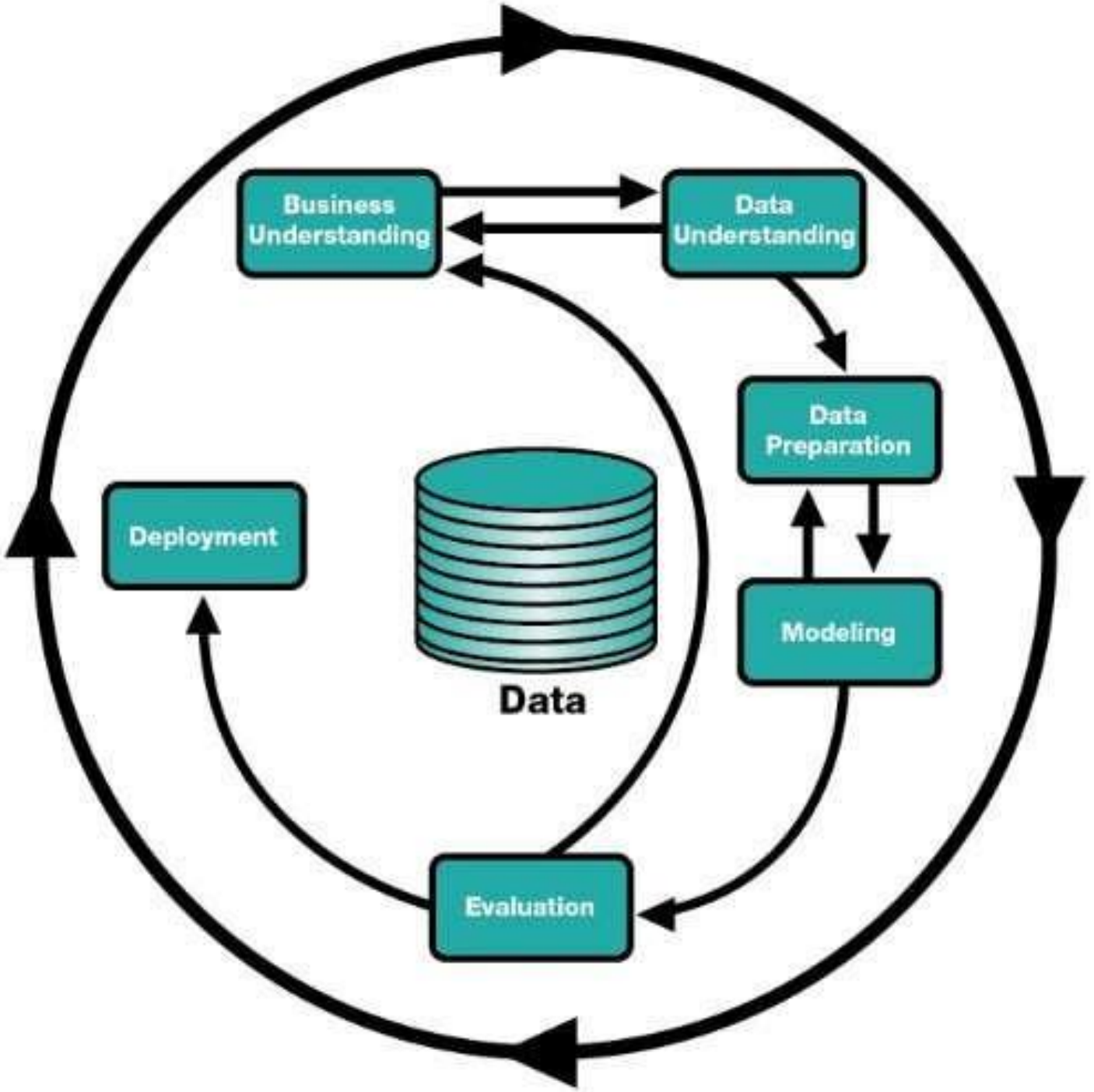
(CRoss-Industry Standard Process for Data Mining)

- It is a data mining process model that describes commonly used approaches that data mining experts use to tackle problems.

CRISP-DM breaks the process of data mining into six major phases:

- BUSINESS UNDERSTANDING: This is the first phase of CRISP-DM process which focuses on and uncovers important factors including success criteria, business and data mining objectives and requirements as well as business terminologies and technical terms.
- DATA UNDERSTANDING: This is the second phase of CRISP-DM process which focuses on data collection, checking quality and exploring of data to get insight of data to form hypotheses for hidden information.
- DATA PREPARATION: This phase focuses on selection and preparation of final data set. This phase may include many tasks records, table and attributes selection as well as cleaning and transformation of data.
- MODELING: This is the fourth phase of CRISP-DM process selection and application of various modeling techniques. Different parameters are set and different models are built for same data mining problem.
- EVALUATION: The process which focuses on evaluation of obtained models and deciding of how to use the results. Interpretation of the model depends upon the algorithm and models can be evaluated to review whether achieves the objectives properly or not.
- DEPLOYMENT:This phase focuses on determining the use of obtain knowledge and results.it also focuses on organizing, reporting and presenting the gained knowledge when needed.

Illustration

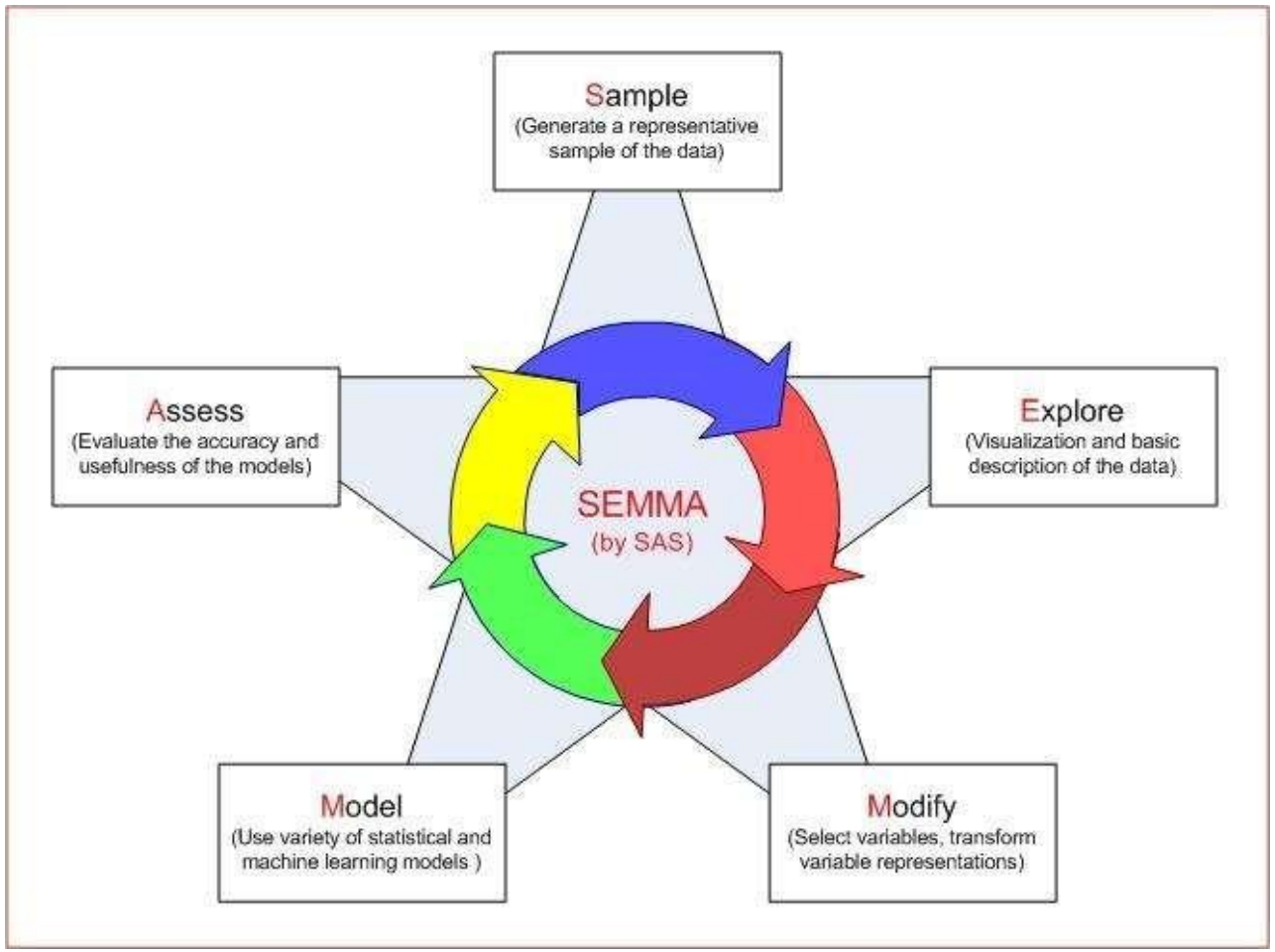


SEMMA PROCESS MODEL

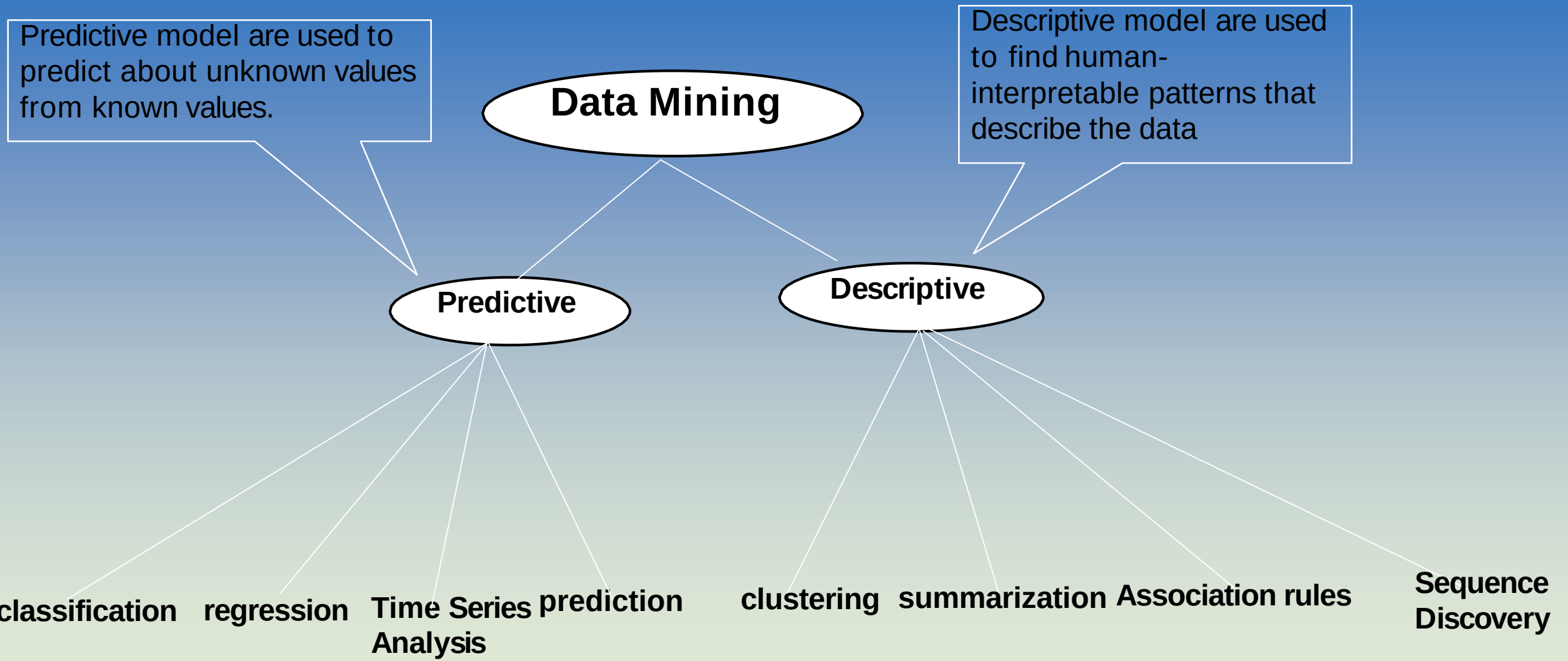
(Sample, Explore, Modify, Model, and Access)

It offers and allows understanding, organization, development and maintenance of data mining projects. It helps in providing the solutions for business problems and goals.

- **Sample:** This stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out as being optional.
- **Explore:** This stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.
- **Modify:** This stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.
- **Model:** This stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
- **Assess:** This stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.


Sample
(Generate a representative sample of the data)
Explore
(Visualization and basic description of the data)
Modify
(Select variables, transform variable representations)
Model
(Use variety of statistical and machine learning models)
Assess
(Evaluate the accuracy and usefulness of the models)
SEMMA
(by SAS)

Data Mining Task



Common Data Mining Tasks

Data mining involves six common classes of tasks Data mining involves six common classes of tasks.

Classification: is the task of generalizing known structure to apply to new data.

Regression: is used to map a data item to a real valued prediction variable.

Clustering: is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data

Association Rules learning : Searches for relationships between variables

Anomaly detection: The identification of unusual data records, that might be interesting or data errors that require further investigation

Summarization: providing a more compact representation of the data set, including visualization and report generation.

1. Classification

Classification

Classification is the process of finding a model that describes the data classes or concepts.

Goal of classification is to build structures from examples of past decisions that can be used to make decisions for unseen cases.

For example, we want to classify an e-mail as "legitimate" or "spam"

CLASSIFICATION: THE PROCESS

CLASSIFICATION: THE PROCESS

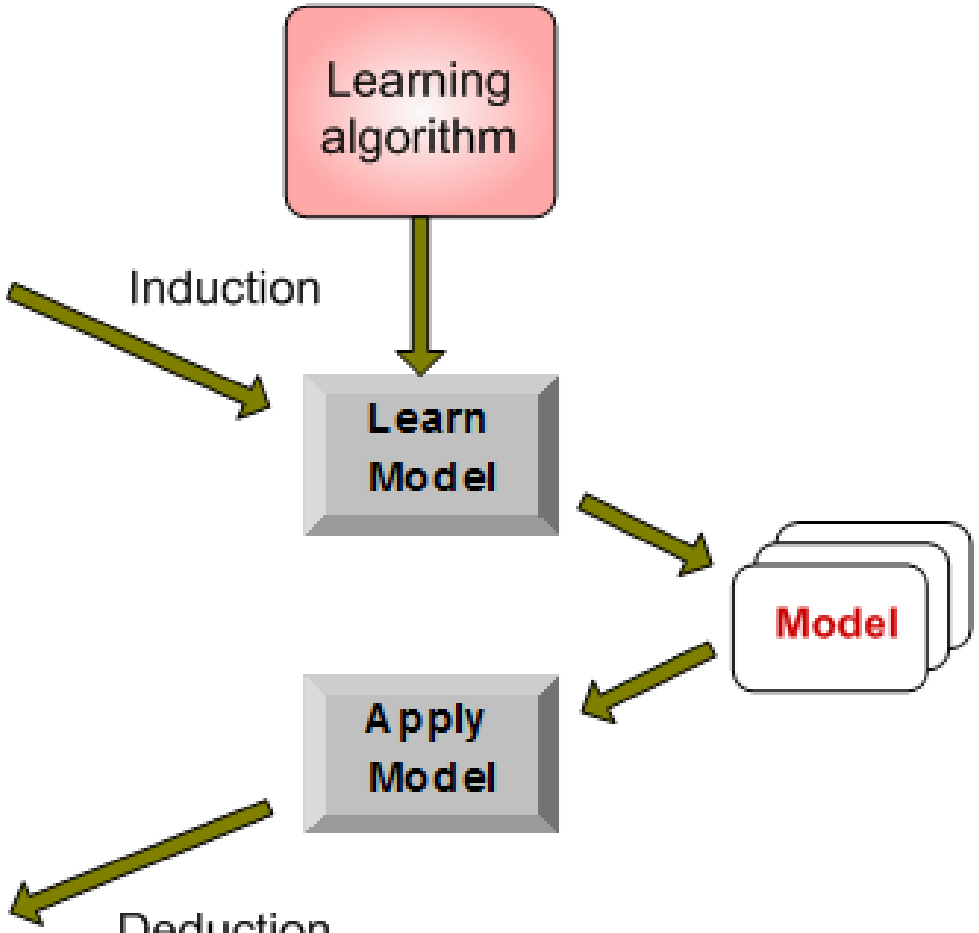
- In classification, we are given a set of labeled examples
- These examples are records/instances in the format (**x,** y) where **x** is a vector and y is the class attribute, commonly a scalar
- The classification task is to build model that maps **x** to y
- Our task is to find a mapping *f* such that f(**x**) = y

Classification: The Process

Classification: The Process

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Data mining software tools

1. RAPID MINER

A data science software platform providing an integrated environment for various stages of data modelling including data preparation, data cleansing, exploratory data analysis, visualization and more.

The techniques that the software helps with are machine learning, deep learning, text mining and predictive analytics.

Easy to use GUI tools that take you through the modelling process. This tool written entirely in Java is an open-source framework and is wildly popular in the data mining world.

2. ORACLE DATA MINING

Oracle, the world leader it database software, combines it's prowess in database technologies with Analytical tools and brings you Oracle Advanced Analytics Database part of the Oracle Enterprise Edition.

It features several data mining algorithms for classification, regressing, prediction, anomaly detection and more. This is proprietary software and is supported by Oracle technical staff in helping your business build a robust data mining infrastructure at the enterprise scale.

3. IBM SPSS MODELER

IBM SPSS Modeller is a visual data science and machine learning solution, helping in shortening the time to value by speeding up operational tasks for data scientists. IBM SPSS Modeler will have you covered from drag and drop data exploration to machine learning.

The software is used in leading enterprises for data preparation, discovery, predictive analytics, model management and deployment. The tool helps organizations to tap into their data assets and applications easily.

4. KNIME

Konstanz Information Miner is an open-source data analysis platform, helping you with build, deployment and scale in no time. The tool aims to help make predictive intelligence accessible to inexperience users. It aims to make the process easy by it is a step by step guide based GUI tools. The product markets itself as an End to End Data Science product, that helps create and production data science using its single easy and intuitive environment.

5. PYTHON

Python is a freely available and open-source language that is known to have a quick learning curve. Combined with is the ability as a general-purpose language and it is a large library of packages that help build a system for creating data models from the scratch, Python makes for a great tool for organizations who want the software they use to be custom built to their specifications.

6. ORANGE

Orange is a machine learning and data science suite, using python scripting and visual programming featuring interactive data analysis and component-based assembly of data mining systems. Orange offers a broader range of features than most other Python-based data mining and machine learning tools. It is a software that has over 15 years of active development and use. Orange also offers a visual programming platform with GUI for interactive data visualization.

7. KAGGLE

The largest community of data scientists and machine learning professionals. Kaggle although started as a platform for machine learning competitions, is now extending its footprint into the public cloud-based data science platform arena. Kaggle now offers code and data that you need for your data science implementations. There are over 50k public datasets and 400k public notebooks that you can use to ramp up your data mining efforts. The huge online community that Kaggle enjoys is your safety net for implementation-specific challenges.

8. RATTLE

The rattle is an R language based GUI tool for data mining requirements. The tool is free and open-source and can be used to get statistical and visual summaries of data, the transformation of data for data models, build supervised and unsupervised machine learning models and compare model performance graphically.

9

. WEKA

Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning tools written in Java. A collection of visualization tools for predictive modelling in a GUI presentation, helping you build your data models and test them, observing the model performances graphically.

10. TERADATA

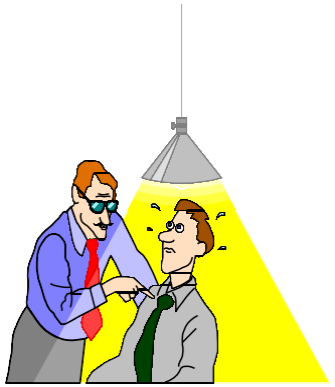
A cloud data analytics platform marketing its no code required tools in a comprehensive package offering enterprise-scale solutions. With Vantage Analyst, you don’t need to be a programmer to code complex machine learning algorithms. A simple GUI based system for

Applications

- **Banking**: loan/credit card approval
  - predict good customers based on old customers

- **Customer relationship management**:
  - identify those who are likely to leave for a competitor.

- **Targeted marketing**:
  - identify likely responders to promotions



- **Fraud detection**: telecommunications, financial transactions
  - from an online stream of event identify fraudulent events

- **Manufacturing and production**:
  - automatically

Applications (continued)

- **Medicine**: disease outcome, effectiveness of treatments
  - analyze patient disease history: find relationship between diseases
- **Molecular/Pharmaceutical**: identify new drugs

- **Scientific data analysis:**
  - identify new galaxies by searching for sub clusters
- **Web site/store design and promotion:**
  - find affinity of visitor to pages and modify layout

Conclusion

- **Data mining**: discovering interesting knowledge from **large** amounts of data .
- **A KDD** process includes data selection , transformation, data mining, pattern